5-2019

Effects of talker variability on categorization of spectrally degraded vowels.

Emily A. Dickey
University of Louisville

Effects of Talker Variability on Categorization of Spectrally Degraded Vowels

By

Emily A. Dickey

Submitted in partial fulfillment of the requirements for

Graduation *magna cum laude*

and

for Graduation with Honors from the

Department of Psychological and Brain Sciences

University of Louisville

March 2019

Abstract

When a person listens to a context sentence with prominent higher frequencies, the subsequent vowel sound is more likely to be perceived as being of a lower frequency and vice versa. This is a spectral contrast effect (SCE). Recent work has shown that talker variability diminishes these SCEs. They were found to be smaller when 200 sentences were spoken by a different talker each time compared to one talker (Assgari & Stilp, 2015). Cochlear Implant (CI) users' speech categorization is also influenced by SCEs but are known to struggle with talker discrimination. Here, I tested whether talker variability affected SCEs when the speech was spectrally degraded. Listeners indicated whether they heard "ih" (as in bit) or "eh" (as in bet) following 200 context sentences spoken by the same talker or varying talkers (Assgari & Stilp, 2015). The sentences were noise vocoded to broadly simulate CI processing. SCEs occurred but did not significantly differ across one-talker and 200-talker conditions. Talker variability does not appear to affect perception in acoustic simulations of CI processing in the same way it does for normal-hearing listeners.

Effects of Talker Variability on Categorization of Spectrally Degraded Vowels

Statistically, about 2 to 3 out of every 1,000 children in the U.S. are born with a detectable level of hearing loss in one or both ears (NIDCD, 2018). Those individuals may use a hearing aid to improve their levels of hearing. The aid amplifies sounds from the environment. However, a person with profound or severe hearing loss would not benefit from a hearing aid. A structure in the inner ear, also referred to as the cochlea, is too damaged to respond normally to sound. For those individuals, a cochlear implant (CI) is a better alternative. A CI is surgically implanted, and electrodes are placed in different locations in the cochlea. The electrical current stimulates any neural fibers that are present. As a result, frequencies are coded (Dorman, 1998). The CI user can then hear sound.

CIs, while allowing patients to hear sounds, do not produce sounds of high quality (Dorman, 1998). Having a CI can allow the person to hear sound, but it is far from a perfect solution. Many people struggle with understanding speech even after receiving a CI depending on the sound quality, how long the person had been deaf before receiving it, and other factors (Dorman, 1998). No type of CI will allow a patient to hear sounds the same way a normal hearing person does.

The main focus of this project is how CIs change how individuals experience spectral contrast effects (SCEs) in hearing. People actually experience contrast effects in perception more commonly than they think but may not be aware of it. For example, when looking at a cell phone screen on a sunny day, it is perceived as being dimmer than when indoors despite the fact the brightness on the screen has not changed. In this example, the sunlight (context) is causing the cell phone screen to be perceived as dimmer (target). In the context of the study that was

conducted, a context sentence affects the perception of the subsequent vowel sound. For instance, after listening to a high-F1 (frequency) context sentence, the subsequent vowel sound will be perceived to be of a lower frequency. In those cases, the participants are more likely to indicate that they heard “ih” as in “bit” because it is a low-F1 vowel sound. The opposite is true as well; after listening to a low-F1 (frequency) context sentence the subsequent vowel sound will be perceived to be of a higher frequency. The participants are then more likely to indicate that they heard “eh” as in “bet” because it is a high-F1 vowel sound. These types of SCEs play an important role in speech perception (Stilp, Anderson, & Winn, 2015).

Recent findings have revealed two important aspects of SCEs in speech perception. First, SCEs occur for CI users (Feng & Oxenham, 2018) and in simulations of CI processing (Stilp, 2017). This indicates that SCEs do not occur only for normal- hearing listeners. Second, SCEs are sensitive to characteristics of who says the context sentence that precedes the target sound. In other words, they are affected by whether the sentences are spoken by the same speaker each time or if each sentence is spoken by a different talker. For example, it has previously been found that the SCEs are smaller when the sentences (as in the example above) were spoken by 200 talkers compared to a single talker (Assgari & Stilp, 2015). This illustrates a difficulty in identifying vowel sounds. More participants should indicate “ih” after hearing a high-F1 context sentence and vice versa. However, smaller SCEs indicate that the frequency of the context sentences had less effect on what vowel sound was perceived and that the participants were less accurate. The explanation for this is the participants became familiar with the single talker but with 200 they were required to constantly readjust to each new talker. The constant recalibration then led to diminished effects in SCEs.

Nonetheless, it is an open question whether CI users (or simulations of CI processing) would exhibit similar sensitivity to talker characteristics because CI users have difficulty discriminating talkers (Fu, Chinchilla, & Galvin, 2004; Massida et al., 2011; Stickney, Zeng, Litovsky, & Assmann, 2004). Cochlear implant users are less sensitive to changes in pitch and vocal tract length than normal-hearing listeners are (Gaudrain & Başkent, 2018), and these are two important cues for identifying a person's sex, age, and size (Smith & Patterson, 2005).

The current experiment specifically measures SCEs with varying talkers in CI processing. On each trial the participants heard a context sentence followed by a vowel sound. Their objective was to label the vowel sound as "ih" or "eh." Each sentence was either spoken by the same talker or by a different talker each time. Despite the fact that the Assgari and Stilp (2015) experiment studied SCEs that occur with varying talkers, for this current study the sentences simulated CI processing.

Two experiments were conducted. In Experiment 1, the spectral resolution, or sound quality, of the CI simulation was either at 4 channels or 8 channels. Channels refer to the pathways that allow information to be sent to the brain and processed (Wilson & Dorman, 2008). Given the extreme difficulty of the task, Experiment 2 was then run where the participant heard each sentence at either 12 or 24 channels. The number of channels was increased for Experiment 2 because the greater the number of channels, the greater quality the sound was (Wilson & Dorman, 2008). Logically, it makes sense that the quality would increase as more channels would mean more pathways and more information being processed.

When predicting the main outcomes of this experiment, it is important to remember what is already known. Examining the results of Assgari and Stilp (2015), SCEs are smaller when

listening to sentences that are spoken by 200 talkers compared to a single talker. This shows that the task is more difficult with a different speaker for each sentence. However, CI users have difficulty discriminating talkers (Dorman, 1998). The participants in the current study and in the study conducted by Assgari and Stilp were normal-hearing listeners but in the current experiment the participants listened to speech that approximated CI processing. It is reasonable to predict that the participants should not be as affected by varying talkers as if the speech was not noise vocoded. Aside from varying talkers the spectral resolution varied as the number of channels varied. Spectral resolution affects the quality of the sound but does affect how well pitch is detected. This indicates that despite hearing the speech at a higher quality, the participants still would not detect pitch as well due to the noise vocoded speech. With this knowledge, the most logical prediction was that there would be little to no interaction between the number of channels and the number of talkers.

Methods

Participants

Forty UofL students signed up for the study on SONA. SONA is an online site where students can sign up to participate in research studies being conducted at the University of Louisville. The participants were awarded with course credit for participating. All participants were required to be at least 18 years old and native English speakers with no known hearing loss. Two studies were conducted. As appearing in the original proposal, the spectral resolution of 4 or 8 channels (n=20) was tested first. The second experiment tested 12 or 24 channels (n=20). All participants completed one of the two experiments, but not both.

Stimuli

In this experiment, the participants heard sentences that simulated CI processing. This was done through noise vocoding the speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). After the full frequency range was divided into a number of channels, the amplitude envelope (changes in volume over time) was then extracted from each channel. The speech-band amplitude envelope was then placed onto white noise with the same frequency bandwidth. Lastly, the frequency bands were combined.

In Experiment 1, the full frequency range was divided into 4 and 8 channels since voice gender identification improves between 4 and 8 channels (Fu et al., 2004; Massida et al., 2011; Stickney et al., 2004). However, even with 4 and 8 channels the identification did not improve enough for the participants to perform well.

This then led to Experiment 2. To test a wide range of signal qualities in this experiment, the full frequency range was then divided into 12 or 24 channels, as in a previous study (Stilp, 2017). A wide range of channels were tested because CI users are highly variable in their performance, so a range was tested to see if spectral resolution had any influence on the interaction between talkers and contrast effects. All of this signal processing was completed in MATLAB with the assistance of a mentor (see Stilp, 2017 for details).

In both experiments, the context sentences and target vowels were the same stimuli from a study conducted by Assgari and Stilp (2015). In the one-talker condition for that experiment, 200 unique sentences were spoken by one male talker (from the HINT database; Nilsson, Soli, & Sullivan, 1994). In the 200-talker condition, each sentence was unique and spoken by a different

talker (from the TIMIT database; Garofolo et al.,1990). These sentences were filtered to amplify low-F1(100-400 Hz) or high-F1 frequencies (550-850 Hz) in order to produce spectral contrast effects (as reported in Assgari & Stilp, 2015). The target vowels were a 10-step series that gradually changed from “ih” as in “bit” to “eh” as in “bet”, spoken by an adult man (See Assgari & Stilp, 2015 for details). Experimental trials consisted of one sentence preceding one target vowel with a silent gap in between that lasted 50 milliseconds.

Procedure

The procedure began with the participant reading and signing the consent form. They were then led into a sound-isolating booth where they put on headphones. A computer then led them through the experiment. On each trial, they heard a sentence followed by a subsequent target vowel. They clicked with their mouse to indicate if they heard that vowel as “ih” as in bit or “eh” as in bet.

Two experiments were conducted, but both were a 2 (number of channels) X 2 (number of talkers) within-subjects design. In both experiments, there were 4 blocks. In Experiment 1 the 4 blocks were: 4 channels with 1 talker, 4 channels with 200 talkers, 8 channels with 1 talker, and 8 channels with 200 talkers. In Experiment 2 the 4 blocks were the same except 12 and 24 channels were tested. They were tested in counterbalanced orders, and each block contained 200 trials in random orders. In the 200 trials, there were 10 target vowels (“ih” and “eh”) multiplied by 2 filtering conditions that either made the sentence be at a higher frequency or a lower frequency multiplied by 10 repetitions. Lastly, every trial was a different sentence. Each participant was allowed a short break after each block. The entire experiment took approximately one hour.

Results

The SCEs were calculated the same way in each experiment. However, the SCEs were calculated in the same way. To produce an SCE, the low-F1-amplified (low frequency) sentences were predicted to produce more "eh" responses (high-F1, or high frequency, responses), and high-F1 amplified sentences were predicted to produce more "ih" (low-F1) responses. The mean percentage of "eh" responses was calculated in low-F1-peak sentences and high F1-peak sentences across all target vowels and repetitions. The difference between these two percentages was calculated, which measured the contrast effect (i.e., how much more often were listeners responding "eh" following low-F1-amplified sentences compared to high-F1-amplified sentences). Similar to this experiment, there was a study conducted previously on contrast effects in cochlear-implant-simulated speech (Stilp, 2017) and the spectral contrast effects for this experiment were calculated using the same approach. SCEs were calculated for each listener for each condition. They were then analyzed using a 2 (number of channels) X 2 (number of talkers) within-subjects ANOVA.

Experiment 1

Experiment 1 tested a spectral resolution of 4 and 8 channels. Following Stilp (2017), if a listener utilized only one response category for more than 80% of trials in the easiest block presented (in this instance, 1 talker at 8 channels), his/her data were removed. This would occur if the participant selected one option more than 80% of the time which illustrates either a lack of ability to discern which vowel sound was heard, or a lack of motivation to complete the task. This resulted in removing 3 participants from the dataset.

According to previous studies (Stilp, et al., 2015) the data showed curves of an “S” shape. However, the main difference is that this current study had participants listen to noise vocoded speech. The line should start near 0 proportion of “eh” responses and end at proportions near 1 but that is not what is seen in the data. When the speech is noise vocoded as was done in previous studies (Stilp, 2017), the S shape was broadly present but only went from approximately 0.2 to 0.8. As can be seen from Figure 1, the task was far more difficult than anticipated. If the listeners are unable to differentiate “ih” from “eh” the presence or absence of a contrast effect cannot be interpreted. The average contrast effect calculated for all participants was 4.33% which shows difficulty in the task as hearing low or high amplified-F1 sentences should not result in the participants choosing the same vowel sounds. There should be a clearer difference in the number of “eh” responses and the number of “ih” responses. An ANOVA was conducted, but the experiment was too difficult to interpret the results.

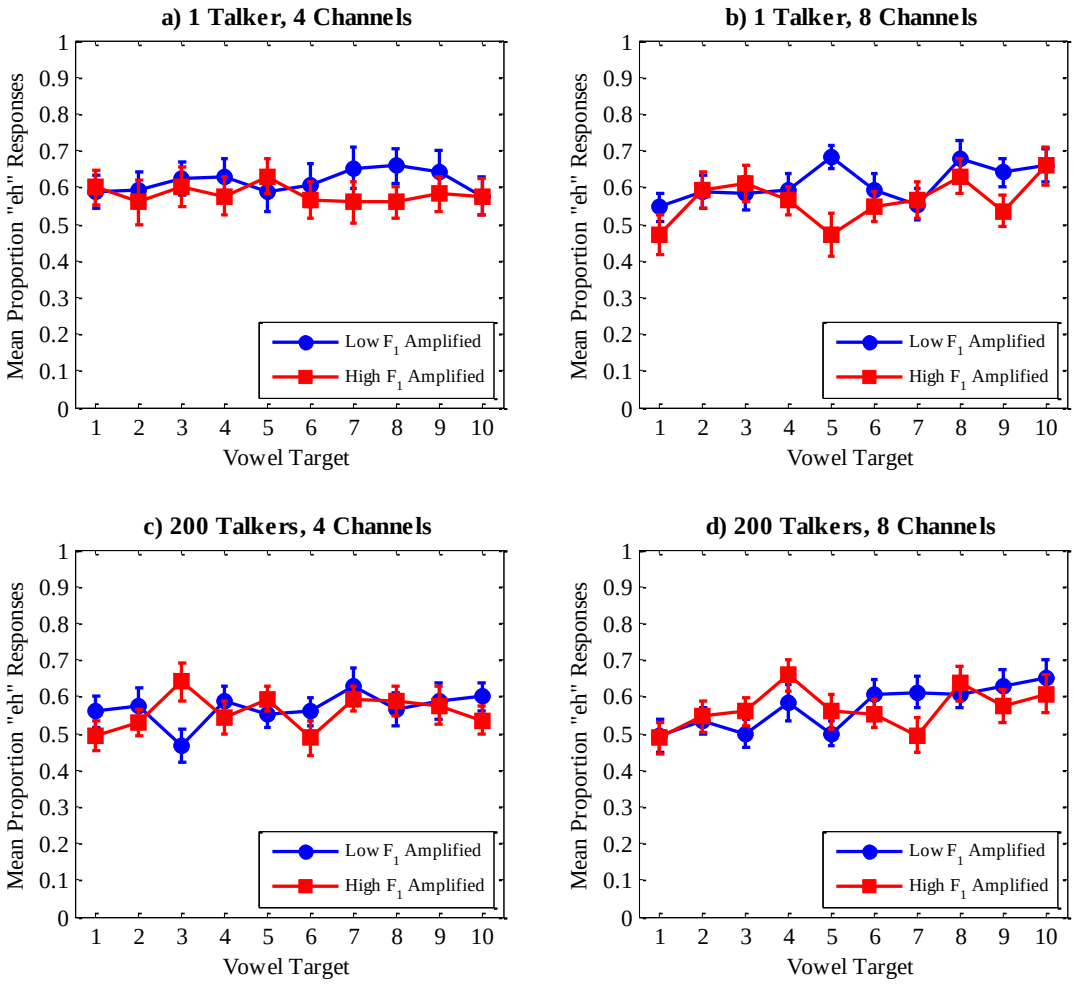

Figure 1: *The y axis represents the mean proportion of "eh" responses to "ih" responses the vowel targets. The circles represent the average number of "eh" responses influenced by context sentences that had lower frequencies (low-F1) amplified, and sentences that had higher frequencies (high-F1) amplified.*

Experiment 2

The results of the first experiment led to conducting Experiment 2. Instead of testing 4 and 8 channels, 12 and 24 channels were tested. This meant there was a higher spectral resolution. The easiest condition was 1 talker, 24 channels and 2 listeners' data were removed for using one response >80% of the time. Just like was anticipated, the higher spectral resolution caused the results to become more S-shaped. This indicated that participants had at least some ability to distinguish "eh" from "ih."

SCEs were calculated for each participant in each block. Three *F* tests were conducted. There was no main effect of number of talkers, $F(1,17) = 0.03$, $p = 0.88$. This can be seen when looking at the top row versus the bottom row where the only difference is the number of talkers. The differences between the SCE percentages are not statistically significant. There was also no main effect of number of spectral channels, $F(1,17) = 0.57$, $p = 0.46$. This is visibly seen when examining the left column versus the right column where the only difference is the number of channels. The differences between the SCE percentages are not statistically significant. Lastly, the interaction between number of channels and number of talkers was not significant either, $F(1,17) = 0.00$, $p = 0.97$. When looking at the percentages of SCEs in all 4 panels, the differences are not statistically significant. Overall, SCEs were indeed present (see gaps between lines in Figure 2 on the following page), but they did not vary based on number of talkers or spectral resolution.

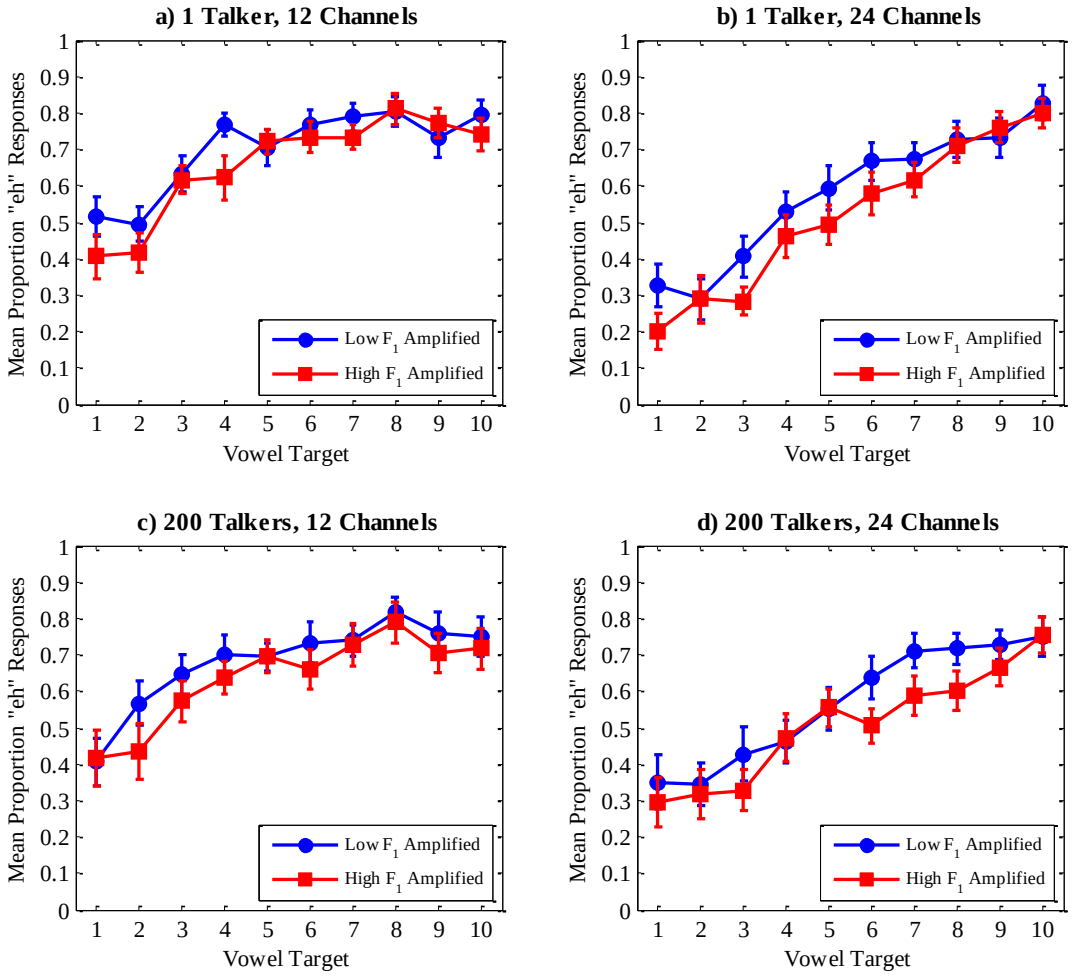

Figure 2: *The y axis represents the mean proportion of "eh" responses to "ih" responses the vowel targets. The circles represent the average number of "eh" responses influenced by context sentences that had lower frequencies (low-F1) amplified, and sentences that had higher frequencies (high-F1) amplified. Twelve and 24 channels are now being tested instead of 4 and 8.*

When the first experiment was conducted testing 4 and 8 channels, the task was too difficult. The participants could not categorize vowels, let alone show contrast effects. When the tested channels were increased to 12 and 24, the participants were better at categorizing, were showing contrast effects, and were not varying based on number of talkers or spectral resolution. The results were consistent with the prediction that the interaction between number of talkers and number of channels would not be statistically significant.

Discussion

The primary objective was to study sensitivity to different talkers in speech that simulates cochlear implant processing. It is already known that CI users struggle with talker discrimination (Fu et al., 2004; Massida et al., 2011; Stickney et al., 2004). It has also been seen with normal-hearing listeners that hearing varying talkers diminishes spectral context effects compared to hearing one talker (Assgari & Stilp, 2015). Here it was tested whether this was also true in cochlear implant simulated processing. The main finding was that spectral context effect did not vary as a function of varying talkers.

Despite the fact that CIs are an alternative to profound or severe hearing loss they do not allow a hearing impaired person to hear sounds clearly. Sounds are of a significantly low quality with a cochlear implant, especially with a lack of pitch. Pitch is a key to distinguishing voices when there are different talkers (Hillenbrand & Clark, 2009). The reason for CI users being less sensitive to changes in pitch than non-CI users is because the electrodes cannot reach all the way to the center when the surgeon inserts the stimulating electrode into the snail-shaped cochlea. It has to stop short. That leaves some neurons without any chance of being stimulated by an electrode, and those neurons encode the lowest frequencies (in the voice pitch frequency ranges). This results in CI users having a limited access to a wide range of frequencies that can be processed). This makes telling voices apart difficult because individuals' voices vary in frequency and if CI users have a limited access to which frequencies are processed, they are unable to differentiate voices.

There are two sides to the argument of whether the disability of CI users to discriminate between talkers is a limitation of receiving a CI. It is indeed difficult for an individual to discern

which of multiple individuals are speaking when in a different room. However, CI users can learn to adapt to these changes. Instead of relying on pitch, they can rely on phrasing or speed of the speech. Just as blind and profoundly deaf people adapt to their surroundings, CI users can learn to adapt as well to reduce the impact of their disability.

This experiment had some limitations. A limitation is that it can become difficult to determine the appropriate spectral resolution to test in the study. For example, the researcher has to determine how many channels the participant will listen to and which channels those are, such as 4, 8, 12 and 24 in this experiment. The problem that occurred in Experiment 1 could occur again if the task is too difficult for the participants. There were no SCEs that could be interpreted because the response functions were flat instead of S-shaped (Figure 1). When Experiment 2 was conducted with a different number of channels, there were SCEs present. The lack of a practice session is another possible limitation. There was a conscious decision not to include a practice. Previous studies did not include one and making participants too comfortable was not desirable in order to make results comparable across studies. Also, too much practice could possibly make participants too good at perceiving noise vocoded stimuli, and that would lead to larger SCEs than those that actually occur. On the other hand, practice trials could assist in interpreting what is occurring where people decline in their performance in the data with 4 and 8 channels.

While the current experiment was not the first experiment to study SCEs with varying talkers or simulations of CIs, it does differ from previous studies. The interactions between the three elements of simulated CI processing, SCEs and varying talkers were examined. In other words, the current experiment did not only test SCEs that occur when listening to varying talkers. It also did not only test SCEs that occur when non CI users listen to spectrally degraded

speech that approximates CI processing. Instead, both aspects were tested in the same experiment.

Moving forward, more experiments should study simulations of CIs using different talkers. The benefit will be to check for consistencies within the data. In a future experiment, the spectral resolution should remain the same as Experiment 2. SCEs were present which showed that with the stimuli, participants showed some ability to differentiate vowel sounds. In addition, the number of participants should be doubled. This would diminish the effects of removing certain participants from the dataset. Lastly, in place of a practice session all participants should take a brief test to screen for undetected levels of hearing loss. It could simply require pressing a button when a certain sound is heard. If the participant does indeed have levels of hearing loss that he or she was unaware of, the researcher can then decide to remove that participant from the dataset. This would be beneficial as the experiment relies on the participants having healthy hearing.

References

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *The Journal of the Acoustical Society of America, 138*(5), 3023-3032. doi: 10.1121/1.4934559

Dorman, M. F. (1998). An overview of cochlear implants. In *Cochlear implants*: *a handbook*, 5-28. Jefferson, NC: McFarland & Company

Fu, Q., Chinchilla, S., & Galvin, J. J. (2004). The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users. *Journal of the Association for Research in Otolaryngology, 5*(3), 253-260. doi:10.1007/s10162-004-4046-1

Feng, L., & Oxenham, A. J. (2018). Effects of spectral resolution on spectral contrast effects in cochlear-implant users. *The Journal of the Acoustical Society of America, 143*(6), EL468-EL473. doi:10.1121/1.5042082

Gaudrain, E., & Başkent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear and Hearing*, *39*(2), 226. doi:10.1097/aud.0000000000000480

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f 0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, *71*(5), 1150-1166. doi:10.3758/app.71.5.1150

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1990). DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, *National Institute of Standards and Technology*, NTIS Order No. PB91-505065. doi:10.6028/nist.ir.4930

Massida, Z., Belin, P., James, C., Rouger, J., Fraysse, B., Barone, P., & Deguine, O. (2011). Voice discrimination in cochlear-implanted deaf subjects. *Hearing Research*, *275*(1-2), 120-129. doi:10.1016/j.heares.2010.12.010

NIDCD. (2018, October 05). *Quick Statistics About Hearing*. (2018, October 05). Retrieved from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085-1099. doi:10.1121/1.408469

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303-304. doi:10.1126/science.270.5234.303

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177-3186. doi:10.1121/1.2047107

Stickney, G. S., Zeng, F. G., Litovsky, R., & Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *The Journal of the Acoustical Society of America*, *116*(2), 1081-1091. doi:10.5772/49992

Stilp, C. E. (2017). Acoustic context alters vowel categorization in perception of noise-vocoded speech. *Journal of the Association for Research in Otolaryngology*, *18*(3), 465-481. doi:10.1007/s10162-017-0615-y

Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. *The Journal of the Acoustical Society of America*, *137*(6), 3466-3476. doi:10.1121/1.4921600

Wilson, B. S., & Dorman, M. F. (2008). Cochlear implants: A remarkable past and a brilliant future. *Hearing Research*, *242*(1), 3-21. doi:10.1016/j.heares.2008.06.005